Comparison of Variant Calling Methods for Whole Genome Sequencing Data in Dairy Cattle

C. F. Baes,[*,†] M. A. Dolezal,[‡] E. Fritz-Waters,[§] J. E. Koltes,[§] B. Bapst,[†] C. Flury,[*]
H. Signer-Hasler,[*] C. Stricker,[#] R. Fernando,[§] F. Schmitz-Hsu,[¦] D. J. Garrick,[§] J. M. Reecy,[§] B. Gredler[†]

[*] Bern University of Applied Sciences, Switzerland, [†] Qualitas AG, Switzerland , [‡]Università degli Studi di Milano, Italy ,
[§]Iowa State University, United States of America, [#] agn Genetics GmbH, Switzerland, [¦] Swissgenetics, Switzerland

ABSTRACT: Accurate identification of SNPs from next-generation sequencing data is crucial for high-quality downstream analysis. Whole genome sequence data of 65 key ancestors of genotyped Swiss dairy populations were available for investigation (24 billion reads, 96.8% mapped to UMD31, 12x coverage). Four publically available variant calling programmes were assessed and different levels of pre-calling handling for each method were tested and compared. SNP concordance was examined with Illumina's BovineHD Genotyping BeadChip®. Depending on variant calling software used, between 16,894,054 and 22,048,382 SNP were identified (multi-sample calling). A total of 14,644,310 SNP were identified by all four variant callers (multi-sample calling). InDel counts ranged from 1,997,791 to 2,857,754; 1,708,649 InDels were identified by all four variant callers. A minimum of pre-calling data handling resulted in the highest non-reference sensitivity and the lowest non-reference discrepancy rates.

Key words: variant calling; next generation sequencing.

INTRODUCTION

The process of translating raw next-generation sequence data into usable variants (single nucleotide polymorphisms (SNPs), short insertions and deletions (InDels), copy number variations (CNVs), etc.) is a specific, sensitive and computationally intensive task. A myriad of alignment (e.g. Bowtie 2 (Langmead and Salzberg (2012)), BWA (Li and Durbin (2009)), Stampy (Lunter and Goodson (2011)), etc.) and variant calling (e.g. the genome analysis toolkit GATK (McKenna et al. (2010)), Platypus (Rimmer et al. (2012)) SAMtools (Li et al. (2009))) software programmes are available. Most of these programmes were designed for human analyses, in which the reference genome is quite good, base coverage is deep and the architecture is relatively well known.

The Bos taurus reference genome UMD3.1 contains ~2.8 billion base pairs, approximately 10% of which are not positioned on any of the 30 chromosomes (Zimin et al. (2009)). Though relatively complete, the UMD3.1 reference genome will not likely allow the same accuracy in alignment, variant calling and further downstream analysis as the human reference. Furthermore, average base coverage in cattle studies is generally lower than that in human NGS studies. Nevertheless, existing software for variant calling in human data can be applied to cattle data, although this is not yet well documented.

It is therefore not yet clear which methods and software works best for variant calling in livestock population structures. Here we systematically compare single and multi-sample calling results achieved using four publically available variant calling software programmes. Additionally, the implications of commonly recommended pre-calling steps are methodically analysed. Through evaluating different variant detection methods, preliminary recommendations for variant calling in dairy cattle are given. Our findings can serve as a reference for choosing variant calling software and whether or not pre-calling steps should be implemented.

MATERIALS AND METHODS

Sample Selection. Sixty-five key ancestors of the main Swiss dairy populations were selected by applying an iterative algorithm, which uses the numerator relationship matrix to rank animals according to percentage of genetic diversity they explain in a given population. Specifically, animals were selected with $p_m = A_m^{-1} \cdot c_m$, where p is a vector that contains the percentage of gene pool diversty captured by m animals selected from the entire genotyped population, A_m^{-1} is a subset of the inverted numerator relationship matrix for m animals and c is a vector representing the average relationship of the m animals selected (Goddard and Hayes (2009)). The subset of selected sires consisted of 34 key Brown Swiss and Original Braunvieh ancestors, that accounted for 74% of the genetic diversity in the genotyped population, as well as 32 key Simmental, Swiss Fleckvieh and (Red) Holstein ancestors that accounted for 74% of the genetic diversity of the genotyped population. Sequence data for these 65 animals was available for analysis (8 Brown Swiss, 18 Braunvieh, 8 Original Braunvieh, 17 (Red) Holstein, 4 Swiss Fleckvieh, and 12 Simmental).

DNA Preparation, sequencing, and alignment. Sequencing was done at the Helmholtz Center in Munich, Germany (German Research Center for Environmental Health Center) in collaboration with the Technical University of Munich. Genomic DNA was extracted from semen samples and sequenced using an Illumina HiSeq2000 (Illumina Inc., San Diego, CA, USA). The bases of the resulting paired-end reads (101 bp), were called with the Illumina BaseCaller, and FASTQ files were produced for downstream sequence data analysis.

Sequence alignment was done according to the sequence alignment guidelines for producing binary alignment mapping (BAM) files for the 1000 bull genomes project (www.1000bullgenomes.com). Briefly, the Burrows-Wheeler aligner (BWA; Version 0.6.1-r104; Li and Durbin (2009)) was used for read alignment to the

Table 1. Number of SNP and insertions and deletions (InDels) found using single and multi-sample calling methods (single and multi-sample calling results include indel realignment and base quality score recalibration)

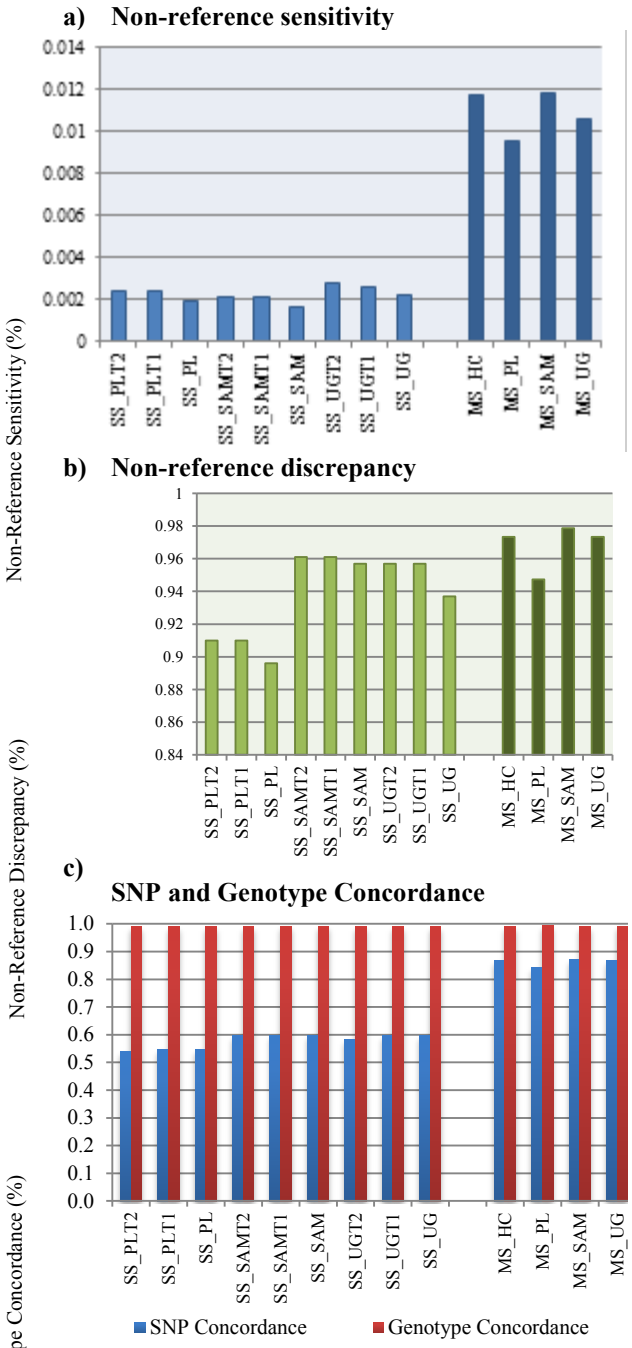
	Total number of SNP identified		Total number of InDels identified	
Caller	**Single Sample**	**Multi-Sample**	**Single Sample**	**Multi-Sample**
Haplotype Caller	-	19,901,885	-	2,685,032
Platypus	17,709,672	16,894,054	2,973,025	2,857,754
Samtools	20,647,891	18,767,273	2,682,094	1,997,791
Unified Genotyper	21,984,283	22,048,382	2,485,677	2,741,468

University of Maryland Bovine reference assembly UMD3.1 (Zimin et al. (2009)). Conversion from sequence alignment map format to sorted, indexed BAM files was done using SAMtools (version 0.1.18; http://sourceforge.net/). PCR-duplicates were flagged using the MarkDuplicates option of the Picard software tools v1.61 (http://picard.sourceforge.net/).

Variant Calling. Both single sample and multi-sample calling methods were applied. Single sample variant calling was performed with three different callers: 1) Samtools MPileup / bcftools (SAM, Li et al (2009)), 2) Platypus (PL, Rimmer et al. (2012)) and 3) the UnifiedGenotyper of the GATK (UG, McKenna et al. (2010)). Three levels of quality recalibration were compared for each animal and caller: a) no quality recalibration (subscript $_{T2}$), b) local realignment around insertions and deletions using the GATK IndelRealigner walker (subscript $_{T1}$, DePristo et al. (2011)) and c) local realignment around insertions and deletions using the GATK IndelRealigner walker followed by base quality score recalibration using the GATK BaseRecalibrator walker (no subscript, DePristo et al. (2011)). Multi-sample variant calling was performed with the same three callers above as well as with the HaplotypeCaller of the GATK (HC, McKenna et al. (2010)). Pre-calling local InDel realignment and base quality score recalibration were conducted for all 4 multi-sample calling analyses.

SNP concordance with Illuminas BovineHD Genotyping BeadChip® was examined. Only Ensemble quality-checked SNPs and SNPs with unique reference sequence (rs) numbers were included in the analysis. Non-reference sensitivity (NRS) and non-reference discrepancy (NRD) (DePristo et al. (2011)) were calculated for one chromosome (BTA24), which we considered representative for the autosomal genome. NRS represents the fraction of polymorphic loci identified by both the caller and the chip over all polymorphic loci identified on the chip. NRS values close to one are desirable. NRD is the fraction of false polymorphic loci identified by both the caller and the chip over all polymorphic loci identified on the chip. Values close to zero are desirable. SNP concordance was calculated as the number of SNPs on the chip identified in the sequence information; genotype concordance was calculated for identified SNP positions as the number of identical genotypes.

Figure 1. Single and Multi-sample caller concordance



SS = Single sample results, MS = Multi-sample calling results, PL = Platypus, SAM = Samtools, UG = UnifiedGenotyper, HC = HaplotypeCaller, T2 = No realignment or recalibration, T1 = local realignment around insertions and deletions using the GATK IndelRealigner walker, (No Subscript) = local realignment around

RESULTS AND DISCUSSION

Descriptive Statistics. Approximately 24 billion paired-end reads were obtained for the 65 sequenced animals. An average of 96.8% of these were mapped to 30 chromosome scaffolds (autosomes 1 – 29, X) of the bovine reference genome assembly UMD3.1 (Zimin et al. (2009)). Approximately 1.7 billion duplicate reads were marked and excluded from further analysis. Average coverage was 12.1 reads per base, with per-animal averages ranging from 10.1

- 17.5. The data have been submitted to the 1000 bull genomes consortium (www.1000bullgenomes.com).

Variant Calling (quality recalibration). Generally, the UG identified the highest number of SNP, followed closely by SAM (Table 1). Different levels of quality recalibration for single sample (SS) calling were examined. For PL and UG, recalibration resulted in slightly fewer SNP, slightly more InDels and slightly fewer multi-allelic sites than when no recalibration was done. For SAM, performing InDel realignment and base quality score recalibration resulted in a slightly higher number of SNPs, InDels and multi-allelic sites identified. These results are in line with those of Liu et al. (2012), who analysed the effect of read pre-processing in whole exome sequencing data and found no pronounced effect of InDel realignment or base quality score recalibration.

Variant Calling (Multi-sample). Variant calling with MS_UG resulted in the highest number of variants; MS_PL had the lowest number of variants (Table1). Le Roex et al. (2012) compared the number of SNP identified with SAM and GATK in African buffalo using ABI SOLiD technology and identified considerably more SNP with GATK. Though not as pronounced, this agrees with both our single sample and multi-sample results.

InDel realignment and base quality score recalibration were conducted before multi-sample calling. As stated by DePristo et al. (2011), multi-sample calling improved NRS for all callers when compared to single sample results (Figure 1a). In contrast, the NRD was slightly poorer in multi-sample calling than in single sample calling, likely because of the inclusion of homozygous reference genotypes in the denominator, which may be difficult to call (Figure 1b). SNP concordance improved considerably through multi-sample calling, due to the inclusion of homozygous reference genotypes, which are not identified in single sample calling (Figure 1c). Genotype concordance was above 99% for all callers and methods (Figure 1c).

CONCLUSIONS

In this analysis, we compared calling methods and preparatory steps for whole genome dairy cattle NGS data. We compared the number of SNP and InDel identified using various publically available variant calling software. Our results show that the number of variants called may differ substantially depending on software. The UG identified the most SNPs in both single and multi-sample calling. We systematically analysed the implications of commonly recommended pre-calling steps such as InDel realignment and base quality realignment. Surprisingly, quality recalibration resulted in lower NRS and NRD with HD Chip information for all callers, although SNP and Genotype concordance improved. Multi-sample calling clearly improved NRS for all callers, but worsened NRD rates. Further analysis must be conducted to ensure only high quality SNP information is used in downstream analysis.

ACKNOWLEDGMENTS

Financial support from the Swiss Commission for Technology and Innovation and the Swiss Cattle Breeders Federation is greatly appreciated.

LITERATURE CITED

1000 Bull Genomes Project (2014). www.1000bullgenomes.com/ Accessed in January 2014.

The 1000 Genomes Project Consortium. (2010). A map of human genome variation from population scale sequencing. Nature 467(7319):1061-1073.

DePristo M, Banks E, Poplin R,. et al. (2011). A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nature Genetics. 43:491-498.

Goddard M.E., Hayes B.J.: (2009). Genomic selection based on dense genotypes inferred from sparse genotypes. Proc. Assoc. Advmt. Anim. Breed. Genet. 18:26–29.

Langmead B, Salzberg S. (2012). Fast gapped-read alignment with Bowtie 2. Nature Methods. 9:357-359.

le Roex N., Noyes H., Brass A. (2012). Novel SNP Discovery in African Buffalo, Syncerus caffer, Using High-Throughput Sequencing. PLoS ONE 7(11):e48792.

Li H. and Durbin R. (2009). Fast and accurate short read alignment with Burrows-Wheeler Transform. Bioinformatics 25:1754-60.

Li H, Handsaker B, Wysoker A, Fennel T et al. (2009). The Sequence Alignment/Map format and SAMtools. Bioinformatics 25:2078-2079.

Liu Q, Guo Y., Li J. et al. (2012). Steps to ensure accuracy in genotype and SNP calling from Illumina Sequencing data. BMC Genomics 13:S8.

Lunter G. and Goodson M. (2011). Stampy: a statistical algorithm for sensitive and fast mapping of Illumina sequence read. Genome Res. 21:936-939.

McKenna A, Hanna M, Banks E, et al. (2010). The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. 20:1297-303.

Rimmer A, Mathieson I, Lunter G, et al. (2012). Platypus: An Integrated Variant Caller (www.well.ox.ac.uk/platypus).

Zimin A., Delcher A., Florea L. et al. (2009). A whole-genome assembly of the domestic cow, Bos Taurus. Genome Biol 10:R42.